New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships

Anubhav Jain[a)]
Energy and Environmental Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

Geoffroy Hautier
Institute of Condensed Matter and Nanosciences (IMCN), Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

Shyue Ping Ong
Department of NanoEngineering, University of California San Diego, La Jolla, California 92093, USA

Kristin Persson
Energy and Environmental Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; and Materials Science and Engineering, University of California Berkeley, Berkeley, California 94720, USA

(Received 15 May 2015; accepted 17 February 2016)

Data mining has revolutionized sectors as diverse as pharmaceutical drug discovery, finance, medicine, and marketing, and has the potential to similarly advance materials science. In this paper, we describe advances in simulation-based materials databases, open-source software tools, and machine learning algorithms that are converging to create new opportunities for materials informatics. We discuss the data mining techniques of exploratory data analysis, clustering, linear models, kernel ridge regression, tree-based regression, and recommendation engines. We present these techniques in the context of several materials application areas, including compound prediction, Li-ion battery design, piezoelectric materials, photocatalysts, and thermoelectric materials. Finally, we demonstrate how new data and tools are making it easier and more accessible than ever to perform data mining through a new analysis that learns trends in the valence and conduction band character of compounds in the Materials Project database using data on over 2500 compounds.

Materials science has traditionally been driven by scientific intuition followed by experimental study. In recent years, theory and computation have provided a secondary avenue for materials property prediction and design. Several successful examples of materials designed in a computer and then realized in the laboratory[1] have now established such methods as a new route for materials discovery and optimization. As computational methods approach maturity, new and complementary techniques based on statistical analysis and machine learning are poised to revolutionize materials science.

While the modern use of the term *materials informatics* dates back only a decade ago,[2] the use of an informatics approach to chemistry and materials science is as old as the periodic table. When Mendeleev grouped together elements by their properties, the electron was yet to be discovered, and the principles of electron configuration and quantum mechanics that underpin chemistry were still many decades away. However, Mendeleev's approach not only resulted in a useful classification but could also make predictions: missing positions in the periodic table indicated potential new elements that were later confirmed experimentally. Mendeleev was also able to spot inaccuracies in atomic weight data of the time. Today, the search for patterns in data remains the goal of materials informatics, although the tools have evolved considerably since Mendeleev's work.

While materials informatics methods are still in their infancy compared to other fields, advancements in materials databases and software in the last decade are rapidly gaining ground. In this paper, we discuss recent developments of materials informatics, concentrating specifically on connecting a material's crystal structure and its composition to its properties (and ignoring, for instance, microstructure and processing). First, we provide a brief history of classic studies based on mining crystallographic databases. Next, we describe the recent introduction of computation-based databases and their

Contributing Editor: Susan B. Sinnott

[a)]Address all correspondence to this author.
e-mail: ajain@lbl.edu

DOI: 10.1557/jmr.2016.80

potential impact on the field, followed by a discussion of techniques and illustrative examples of modern materials informatics. Finally, we present a novel materials informatics study, based exclusively on openly-available data sets and tools, that predicts the valence and conduction band character of new materials. We note that while this review is mainly focused on periodic solids, molecular systems have also been extensively studied through data mining approaches.[3–5]

I. EARLY EXAMPLES OF DATA MINING CRYSTALLOGRAPHIC DATABASES

The earliest and still today the most systematic and organized data sets in materials science are based on crystallographic data. Crystal structures of observed compounds are available in databases such as the Inorganic Crystal Database (ICSD),[6] the Cambridge Structural Databases,[7] and the Pauling file.[8] Information on unit cells, atomic positions, and symmetry are available from these resources for hundreds of thousands of inorganic compounds.

Crystal structure data have been extensively used to perform, before the term even existed, data mining studies. For instance, ionic radii for the elements were extracted from large crystallographic data sets in the beginning of the 1970s by Shannon.[9] This was followed by a more complex description of bonding through the bond valence formalism,[10,11] which relates the valence of cation i to a sum of the bond strengths s_{ij} between cation i and anions j through an analytical expression, i.e., $V_i = \sum_j s_{ij}$ in which the s_{ij} terms are summed over bonds through a simple mathematical expression such as $s_{ij} = (r_{ij}/r_0)^{-N}$ or $s_{ij} = e^{\frac{(r_0 - r_{ij})}{B}}$. The parameters r_0, N, or B are specific to a cation–anion pair (ij) and must be extracted from a data set through a fitting procedure. Using the ICSD database,[6] Brown et al. extracted these parameters from 750 atom pairs, building a bond valence table that remains widely used today.[12]

More specific questions on bonding in solids have also been answered through the usage of large structural data sets, e.g., the nature of hydrogen bonds[13] and bonding in borates.[14] Furthermore, these data sets have enabled studies of the distribution of inorganic compounds among space groups,[15,16] to the search for materials with specific crystallographic and symmetry requirements such as ferroelectrics,[17] and for screening structure-based properties such as diffusion paths.[18]

Another early application of crystallographic databases was in crystal structure prediction. Early approaches used structure maps, which plot the experimentally observed type of crystal structures against intuitive chemical descriptors (electronegativity, Mendeleev number, or ionic radius).[19–22] The groupings that form on these maps can then be used to infer the structure in which a new chemical compound will crystallize. An example of a structure map is plotted in Fig. 1 for the A_1 B_1 stoichiometry. Recently, Morgan et al. used modern cross-validation techniques to demonstrate that structure maps are indeed predictive and quantified their performances.[23] Structure maps have also been combined with modern data mining techniques to build a predictive model using information entropy and classification trees for the prediction of binary halide scintillators.[24]

II. NEW RESOURCES: THE ADVENT OF COMPUTATIONAL MATERIALS DATA REPOSITORIES AND OPEN SOFTWARE

While it is possible to perform data mining using crystal structures alone, most informatics studies additionally require materials property measurements. Although many databases of experimental materials properties are now available, it can be difficult to extract large-scale structure-property relationships from these resources. Computational databases, while also possessing many important limitations, may be able to supplement the capabilities of experimental databases and facilitate an informatics-style approach to materials design.

A. Experimental materials databases

One class of experimental materials databases are the crystallographic structure repositories mentioned previously, which include the ICSD,[6] the Pauling file,[8] CRYSTMET,[25] and Pearson's Crystal Data.[26] These resources have been recently summarized and reviewed by Glasser.[27] Materials properties databases are also available. The largest of these is likely the set of databases from Springer, which includes the comprehensive Landolt–Börnstein Database.[28] However, most materials property information remains scattered across multiple resources, including the FactSage suite of databases,[29] the National Institute of Standards and Technology databases,[30] MatWeb,[31] MatNavi,[32] various publications such as the Kubaschewski tables,[33] and the Handbook of Ternary Alloy Phase Diagrams.[34] We note that Citrine Informatics (http://www.citrine.io) is one commercial entity that is attempting to centralize information collected from diverse sources (both experimental and computational).

These various data sources have historically been expertly curated and validated, and serve as important, trusted resources for the materials research community. While it is certainly possible to perform data mining on these databases, limitations include **completeness** and **programmatic access**. In terms of completeness, many materials properties (e.g., formation energies, band gaps,

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

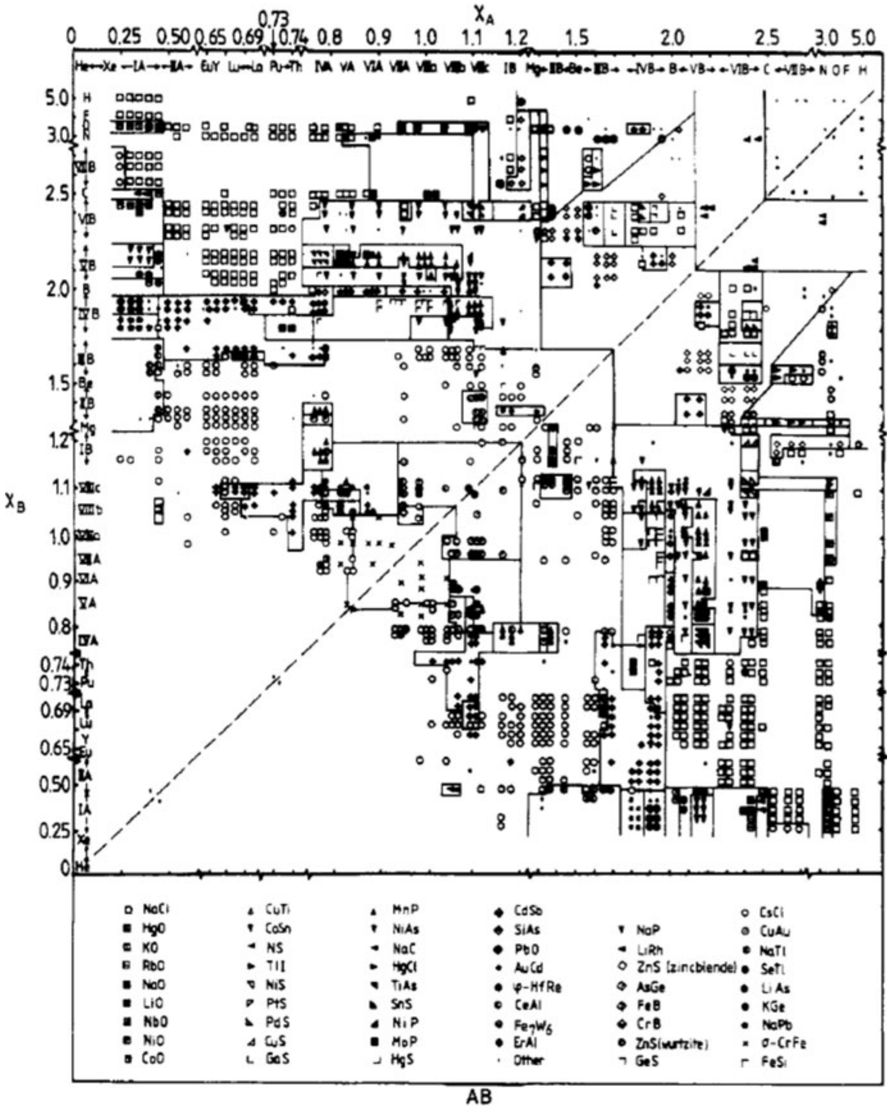

FIG. 1. An example of a structure map for the A_1 B_1 composition. Each symbol indicates a specific crystal structure prototype. The axis refers to a “chemical scale” attributing a number to each element based on its position in the periodic table (Mendeleev number). Image from Ref. 20. ©IOP publishing, all rights reserved. Reprinted with permission.

and elastic tensors) have only been measured for only a small fraction of the number of known crystal structures. There is particularly a lack of data available for negative results, including failed synthesis attempts and unexceptional materials properties measurements. Even when compound properties are available, they are often associated only with a composition and lack a rigorous description of the material being measured (crystal structure, microstructure, doping level, etc.). The lack of information about the input material can make it very challenging to develop models. Finally, in terms of data access, most of these databases can only be accessed through mechanisms designed for “single lookups” rather than systematic data mining over large portions of the database. Thus, there is room for other types of databases that can help address the gaps in the experimental record.

B. Computational materials databases

In recent years, the ability to *generate* materials data using systematic high-throughput computations (typically based on density functional theory, or DFT, approaches for solving the Schrödinger equation[35,36]) has created new, efficient opportunities to produce high-quality databases for data mining. These computationally-driven databases, which usually leverage crystal structure information from experimental databases, provide powerful means to extract patterns and correlations from hitherto unavailable data sets. As an example, the full elastic tensor has only been measured for approximately 150 distinct compounds, but a recent high-throughput computational study has tabulated this quantity for over 1000 materials.[37]

Examples of such computationally-derived databases include the Materials Project,[38] AFLOWlib,[39] the Open Quantum Materials Database,[40] the Harvard Clean Energy Project,[41] the Electronic Structure Project,[42] NoMaD,[43] NRELMatDB,[44] and the Computational Materials Repository.[45] Some of these databases can be quite extensive; for example, the Materials Project[38] today contains property data for over 60,000 compounds and includes many different properties, and AFLOWlib[39] includes over 600,000 entries. However, more focused efforts are also proliferating, including CatApp[46] for catalysis, PhononDB[47,48] for phonons, TEDesignLab for thermoelectrics,[49] and ESTEST[50] for verification and validation of physics software. In some cases, the separation between these efforts is clear. For example, the Harvard Clean Energy Project[41] is geared toward small molecules, whereas AFLOWlib[39] targets inorganic compounds. In other cases, such as for the Materials Project,[38] AFLOWlib,[39] and the Open Quantum Materials Database,[40] there is considerable overlap in the intended scope. Even in this latter case, users can still benefit from multiple databases, e.g., to verify results or to "fill in the gaps" of their preferred database. Unfortunately, at present there exists no search-engine or similar tool to facilitate search across databases (e.g., something in the spirit of ChemSpider[51]). This might partly be due to the current difficulty of accessing data programmatically in many of these resources, as discussed in Sec. II. C. A summary and comparison of these different efforts can be found in a recent review by Lin.[52]

A major contributing factor to the rise of simulation-based data is the availability of software libraries that have brought large-scale data generation and data mining within the reach of a greater number of research groups. Examples include pymatgen[53] (materials analysis, plotting, and I/O to DFT software), ASE[54] (structure manipulation and DFT calculator interface), AFLOW[55] (high-throughput DFT framework), AiiDA[56] (workflow management for high-throughput DFT), and FireWorks[57] (general workflow software for high-throughput computing). These codebases, as well as the continually improving accuracy of theoretical techniques, robust and more powerful DFT software, and the exponential growth of computing power will likely make simulation-based data sets even more valuable and prevalent in the future.

C. Programmatic data access

An efficient method to download large data sets from data resources (whether experimental or computational) is necessary for performing materials informatics. There are many methods by which data can be exposed, including direct download of either raw or processed data sets. A more modern technique to expose a data resource is to use representational state transfer (REST) principles to create an application programming interface (API) to the database.[58] This method was pioneered in the computer science community and was introduced to the materials world through the materials API (MAPI)[59] of the Materials Project. To date, the MAPI has served more than 15 million pieces of materials data for over 300 distinct users, enabling new types of applications and analyses.

Under RESTful design, each object is represented as a unique resource identifier (URI) that can be queried in a uniform manner using the hypertext transfer protocol (HTTP). Each document or object (such as a computational task, crystal structure, or material property) is represented by a unique URI (see Fig. 2 for an example) and an HTTP verb that can act on that object. In most cases, this action returns structured data that represents the object, e.g., in the javascript object notation format (JSON).

Some of the advantages of RESTful interfaces include:

(i) Abstraction: RESTful interfaces use universal protocols that can be accessed by many programming languages. They hide the details of the underlying data storage implementation (i.e., whether the data is stored in an SQL or NoSQL database), by exposing a clean and consistent set of actions and queries that can be performed against the data.

(ii) Flexibility: Because they abstract away implementation details, RESTful interfaces are flexible to changes in the underlying infrastructure. They also allow for federation amongst several databases with different internal architectures under a consistent API, such that a user can in principle write the same code for different resources. Such flexibility might become especially important in building universal access modes to disparate data sources.

(iii) Power: High-level interfaces can be built upon RESTful APIs such that one can access and manipulate offsite data resources in an object-oriented way. For example, a high-level interface to the MAPI[59] is provided in the pymatgen[53] code base and allows users to obtain

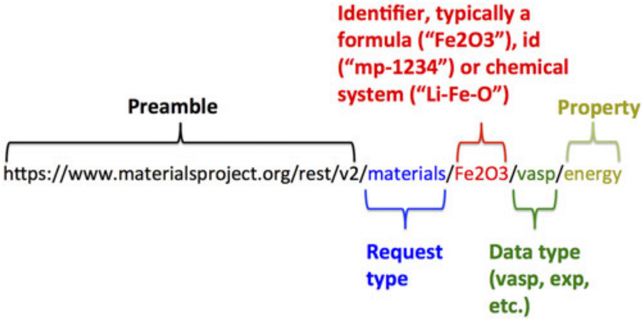
Identifier, typically a formula ("Fe2O3"), id ("mp-1234") or chemical system ("Li-Fe-O")
Preamble
Property
https://www.materialsproject.org/rest/v2/materials/Fe2O3/vasp/energy
Request type
Data type (vasp, exp, etc.)

FIG. 2. An example of the URL structure of the MAPI. Figure reprinted from Ref. 59 with permission from Elsevier.

properties, such as crystal structure or electronic band structure, using built-in functions rather than by explicitly making HTTP requests.

(iv) Up-to-date: Data sets can become stale and outdated. A RESTful interface allows for the most recent version of the data and queries to be exposed at all times without actions needed by the user. Users can always choose to retain data, and the URI scheme can also be used to preserve multiple versions of the data. However, RESTful APIs make it simple to obtain the most current and relevant data for a given analysis without needing to re-download the entire database.

Although REST interfaces can be tricky for beginning users, a well-designed REST interface promotes discoverability of the data and frees the end user from learning the implementation details of a particular database, instead allowing data analysis through a clean and consistent API.

III. MODERN DATA MINING TECHNIQUES AND EXAMPLES

With the generation of ever-expanding materials data sets underway, the major remaining challenges are to develop descriptors (sometimes called “features” or “predictors”) of materials and relate them to measured properties (sometimes called “outputs” or “responses”) through an appropriate data mining algorithm. In the last few decades, many new approaches have been developed to extract knowledge from large data sets using elaborate mathematical algorithms, leading to the new field of machine learning or data mining.[60] In many cases, such algorithms can be applied in an “off-the-shelf” way for materials problems; in other cases, materials scientists have themselves developed new approaches for data analysis that are tuned to their domain.

A. Descriptors for materials structure and properties

Because data mining operates on numerical data structures, materials scientists must first encode materials in a format that is amenable for finding relationships in the data. While several data formats have been developed for describing crystallographic materials (e.g., the CIF file format), these formats are not suitable as data mining descriptors for reasons that follow. The problem of developing robust descriptors for crystalline solids remains a challenging task; here, we identify and recapitulate[61] four properties that characterize good descriptors:

(i) Descriptors should be **meaningful**, such that relationships between descriptors and responses are not overly complex. For example, whereas the lattice vectors and atom positions of a crystal structure in principle determines its properties, such an encoding involves a very complex relationship between inputs and outputs (i.e., the Schrödinger equation). In particular, the complex and important three-dimensional boundary conditions implicit in this representation are not captured by today’s data mining techniques. Better descriptors have simpler relationships to the outputs, ideally within the complexity space of what a data mining algorithm can reasonably uncover (in accordance with similar principles outlined by Ghiringhelli et al.[61])

(ii) Better descriptors are **universal**, such that they can be applied to any existing or hypothetical material. While this is not strictly necessary when performing analysis within a well-constrained chemical space, it is useful for building universal models that bridge chemistries and structures.

(iii) Better descriptors are **reversible**, such that a list of descriptors can in principle be reversed back into a description of the material. This is not strictly necessary for a successful model, but it would enable more efficient “inverse design” in descriptor space rather than in the space of materials. A less stringent version of this condition was put forth by Ghiringhelli et al.,[61] who stated that descriptors should uniquely characterize a material.

(iv) Descriptors should be **readily available**, i.e., they should be easier to obtain than the target property being predicted.[61]

It is unlikely that any one set of descriptors will meet such criteria across the space of all potential compositions, crystal structures, and targeted output properties. Rather, descriptors will likely need to be tailored to an application, as was demonstrated by Yang et al. for topological insulators.[62]

A common first-level set of descriptors is to encode compounds as a vector that depends only on the identity of the elements contained within the compound, without explicit consideration of crystal structure or stoichiometry. For example, an A–B binary compound might include properties such as the electronegativity of the pure A element, the electronegativity of the pure B element, the atomic radii of these elements, and the experimentally known melting points of these elements. One weakness of this strategy is that the properties of pure elements do not always correlate well to their properties exhibited in compounds; as an extreme example, the elemental properties of oxygen gas are very different from the anionic properties of oxygen in oxide compounds. An extension of this basic technique, which we refer to as the “atoms-in-compounds” approach, still only considers the identity of the atoms in a compound when formulating descriptors, but uses descriptions of the elements from known *compound* data. For example, for the same A–B compound we might include descriptors such as the ionic radius of A (determined from its bonds

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

in the context of selected compounds) or the oxide heat of formation of B. Depending on the type of compound being modeled, such descriptors might be more predictive than properties of the pure elements.

Important research efforts are being directed toward the development of novel descriptors today for encoding more complex materials data such as crystal structure. For instance, beyond the description of materials by space groups, a description by crystal structure prototypes can be essential and several algorithms have been developed[63,64] and applied toward probabilistic crystal structure prediction models.[65,66] Further methods include the development of the "symmetry functions" that capture two and three-body terms[67] and local environment descriptors[68] to characterize bonding details in crystals. "Fingerprints" representing the somewhat complex object of crystal structures have also been proposed to characterize the landscape of possible structures and as descriptors in data mining approaches.[69,70] Finally, other recent approaches are inspired by molecular data mining; the idea of a "coulomb matrix" as a representation of organic molecules has been extended to solids by Faber et al.,[71] and the SimRS model for structure primitives has been applied to solids by Isayev et al.[70]

In addition to descriptors derived from experimental data and structural analysis, descriptors from DFT computations are becoming increasingly popular. For instance, a data mining approach using linear regression recently revealed that information from the charge density of the material could be predictive of its elastic tensor.[72] Mixing of computed and traditional descriptors is also possible; for example, Seko et al.[73] presented results on the development of a linear regression model predicting melting temperature from both classical descriptors and computed quantities (such as cohesive energies). This study highlights the gain in predictive power provided by the addition of these computed quantities (see Fig. 3), provided that performing the computations remains more convenient than determining the final property. We note that even properties for which the underlying physics is unclear, such as high-T_c superconductors, can sometimes be tackled by combining computed data (density of states (DOS) and band structures) with experimental observations.[70]

Often, such fundamental descriptors perform better if they are combined through functions; for example, a descriptor describing the *ratio* between the ionic radius of elements A and B in a compound may produce more accurate linear models than two separate descriptors describing the individual radii. We note that functions that are symmetric to the interchange of elements can be used to avoid the problem of site differentiation across crystal structures. One problem in formulating such functions, and with descriptor selection in general, is the large number of possibilities and the danger of using correlated predictors, and in over-fitting. An active area of research in statistical learning methods is in such *feature selection* problems, which identify the best descriptors within a large pool. For example, recent work by Ghiringelli et al.[61] have demonstrated that physically relevant models can be constructed by mathematically selecting between approximately 10,000 possible functional forms of descriptor combinations. Such techniques to automatically derive physical relations are reminiscent of earlier work in physics data mining, in which the equations of motion of a double pendulum could be automatically derived given explicit input variables and an output data set.[74]

B. Exploratory data analysis and statistics

Once the descriptors and target outputs are determined, an appropriate data mining scheme must be selected to relate these quantities. The first step is to plot visual correlations and apply standard statistical tools (such as

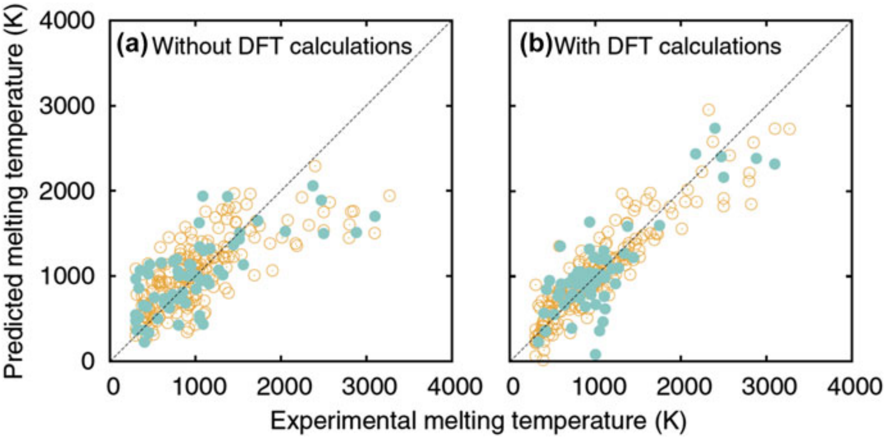

FIG. 3. Predicted versus experimental melting temperature over a data set of 248 compounds for two models: one including DFT descriptors and one without. Image from Ref. 73. Reprinted with permission, ©American Physical Society.

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

the analysis of variance, or ANOVA, approach) to better understand the data set and to make basic predictions. This phase is often referred to as *exploratory data analysis*. For example, one large *ab initio* data set[75] targeted the discovery of new Li-ion cathode materials. This data set encompassed tens of thousands of materials, which were extensively analyzed for trends, limits, and opportunities across cathode chemistries. For example, statistics for this data set were compiled for phosphate materials,[76] which have been of great interest for the battery community due to the success of $LiFePO_4$ as a Li-ion cathode. Using more than 4000 lithium-containing phosphate compounds, the expected voltage for all potential redox couples in any phosphate cathode material could be derived and rationalized. Furthermore, several hypotheses, proposed in the literature and based on limited data, on the relevant factors determining voltage (such as ratio of P/O, number of linkages between phosphate groups and redox metals, and P–O bond length) were confirmed or challenged.

Visual exploration can also uncover relationships between properties. The same project mentioned above uncovered a correlation between the computed average voltage and thermodynamic chemical potential of O_2 (translated into a temperature for oxygen gas release) for cathode compounds (Fig. 4).[77,78] This analysis determined that higher voltage compounds are in general at greater risk for thermal instability. Although a similar idea was proposed much earlier,[79] the advent of high-throughput computed data tested the idea on a large scale and further uncovered chemistry-based differences in behavior.

C. Clustering

Going beyond visual examination and basic correlation analysis, one can use **clustering** techniques to uncover hidden relationships in the data. Clustering divides data into groups based on a similarity metric in a way that uncovers patterns and categories, but does not directly predict new values. Several algorithms exist today for automatically clustering data (e.g., *k*-means clustering, Ward hierarchical clustering, Density-based spatial clustering of applications with noise (DBSCAN), and others[80,81]). To perform a clustering analysis, one must be able to define a distance metric between data points (e.g., Manhattan distance or Euclidean distance between feature vectors) and sometimes additional parameters depending on the algorithm (e.g., a standard *k*-means clustering requires setting the number of clusters in advance); several different trials of an iterative algorithm may be needed to find an optimal grouping.

The results of a clustering analysis can be used downstream in data analysis; for example, each cluster might represent a "category" of material, and a separate predictive model can be fit for each cluster. For example, this is the approach taken in the cluster-rank-model method developed by Meredig and Wolverton,[82] which was used to classify and then predict the stability of various dopants into zirconia.

One technique for visualizing a hierarchical clustering process is the *dendrogram*, which displays the results of clustering as a function of cutoff distance. An example of a dendrogram for two dimensions of clustering is depicted in Fig. 5. This diagram was created by Castelli and Jacobsen[83] to understand which elements can be placed in the A and B sites of a perovskite structure to produce a stable compound that also has a nonzero band gap, with the end goal being to understand principles for designing photocatalyst materials. The researchers observed several "pockets" representing combinations of

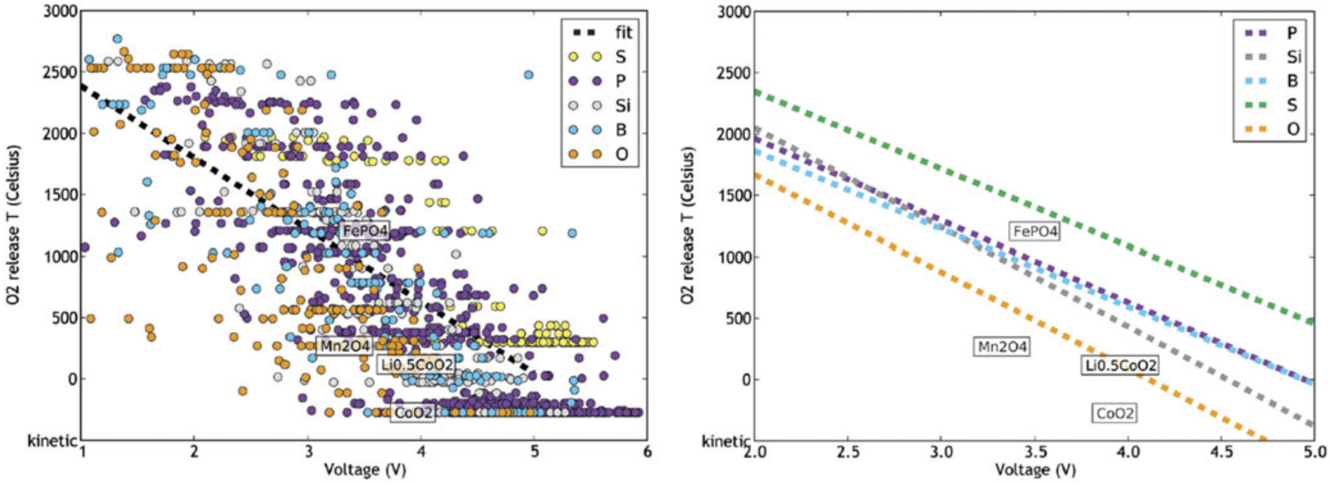

FIG. 4. Temperature of O_2 release versus voltage for a large set of cathode materials. Higher temperatures are associated with greater "safety" of the cathode material. A clear correlation between higher voltage and lower temperatures for releasing oxygen gas is observed. The figure on the right depicts a linear least-squares regression fit to the data for different chemistries (oxides, sulfates, borates, etc.). While all cathode materials have a similar tendency to be less safe for higher voltage, there is a clear difference between different (poly)anions. Image from Ref. 77. Reproduced by permission of the PCCP Owner Societies.

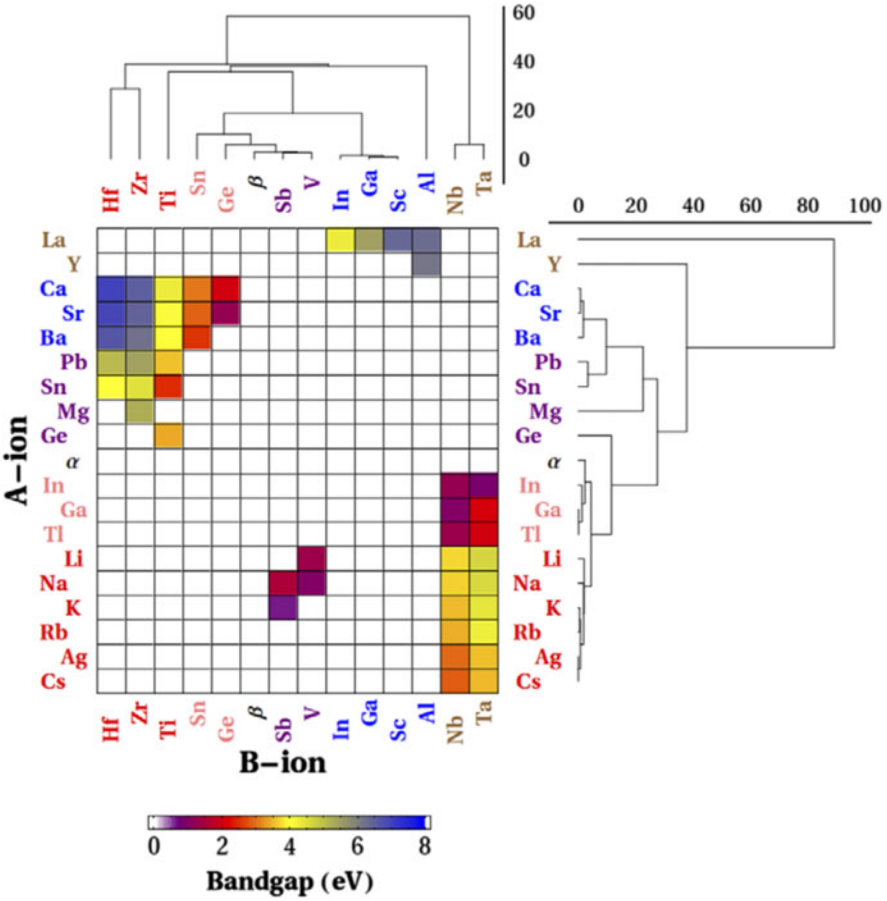

FIG. 5. Elements forming stable oxide perovskites in the A and B sites; the gap is represented by the color, and elements are ordered by size of gap and grouped by similarity in gap. The dendrogram trees for A and B sites are plotted at the right and top of the image, respectively. Image from Ref. 83. ©IOP publishing, all rights reserved. Reprinted with permission.

elements that met their criteria (Fig. 5) and that could be rationalized through the periodic table. The way to read the dendrogram tree in Fig. 5 is as follows:

(i) The dendrogram can be cut at any distance; the number of lines intersecting that cut is the number of clusters formed at that distance cutoff.

(ii) As one progressed down the tree (reduces the distance cutoff), more clusters emerge representing groups of greater homogeneity. For example, in terms of the A ion (right side of Fig. 5), the first "cut" separates La from the other elements, and the second cut separates Y; further cuts divide the remaining elements into groups. In practice, such hierarchical clustering algorithms can either be agglomerative (start with each data point as a separate cluster and begin merging them) or divisive (start with all data points in a single cluster, and start dividing those clusters).

(iii) By coloring the leaf nodes, the distinct clusters at a selected cutoff can be emphasized.

Clustering is often combined with methods for dimension reduction such as principal components analysis, which determines orthogonal directions in high-dimensional space that explain the most variance in the data. Often, only a few principal components are needed to explain most of the data variance, allowing clustering to be performed in a small number of dimensions (e.g., 2 or 3) and subsequently visualized. One example of such a clustering analysis is the classification of oxide DOS spectra performed by Broderick et al.[84] In this study, DFT-based DOS data for 13 compounds were normalized and aligned such that each DOS was parametrized as a 1000 element vector relating energies to number of states. A principal components analysis was then applied to determine which parts of the DOS explained the greatest difference between materials; i.e., the first principle component, which is itself a 1000-element vector resembling a DOS, was highly related to the average difference in the DOS spectrum between monoclinic and cubic/tetragonal structures. Each DOS in the data set was then projected in the space of the first two principle components, and a clustering analysis was performed to find similarities in this space. By examining the clusters, it was determined that variation in the (normalized, aligned) DOS was most affected in the order structure >stoichiometry >chemistry, although it should be noted that there exist hidden relationships between these variables in the data.

A larger-scale version of DOS classification was performed by Isayev et al.,[70] who grouped together structures in the AFLOWlib[39] database through the use of a "D-fingerprint" to encode DOS information (similar to the encoding used by Broderick et al.[84] described previously). Rather than assessing similarity in the space of principal components, the distance between these D-fingerprint vectors was computed directly and used to construct force-directed graphs that placed materials with similar DOS together. While the resulting data set is complex, some of the major groups that emerged were those with a similar number of distinct elements in the compound as well as a similar number of atoms in the unit cell. The authors of the study used such graphs only for making qualitative statements, and explored other methods for performing quantitative analysis.

D. Linear models

When quantitative predictions are required, one option is to construct linear models, in which the response value is calculated as some linear combination of descriptor values. Linear models encompass a broad range of techniques and can often be modified to tackle different types of problems. For example, when predicting probabilities $Pr(X)$ with possible values strictly in the interval (0,1), a limitation of standard linear models is that values outside the interval can be predicted. However, by transforming probabilities to a "log-odds" formalism ($\log[Pr(X)/(1-Pr(X))]$) in *logistic regression*, the interval of prediction can be reformulated into $(-\infty, +\infty)$. If factor variables (i.e., categorical variables such as {"metal", "semiconductor", "insulator"}) are involved, linear models can be used by encoding the potential factor levels as a set of binary coefficients. As a final example, "robust linear models" are resistant to outliers and violations of least-squares linear model assumptions (such as the presence of heteroscedasticity).[85]

Linear methods can be powerful methods for estimation and prediction, particularly if descriptors are well-chosen. For example, Chelikowsky and Anderson[86] determined that the melting points of 500 intermetallic alloys was highly correlated with a simple average of their (known) unary end-member melting points; however, other descriptors (such as differences in Pauling electronegativity) yielded only weak correlations.

One strategy to find good descriptors for linear models is to begin by including many possibilities and subsequently filter out those that are nonpredictive or redundant. For example, the method of partial least-squares fitting can reduce the effect of collinear descriptors (e.g., components of spectra in which neighboring data points are likely to be highly related). Another technique is regularization of the linear coefficients (i.e., reducing the l_1 and/or l_2 norm) through approaches such as least absolute shrinkage and selection operator (LASSO)[87] (penalizes l_1 norm), ridge regression (penalizes l_2 norm), or Elastic net[88] (penalizes a combination of l_1 and l_2 norm). These methodologies help reduce the number and strength of correlated or unhelpful predictors and have been successfully applied to several materials predictions problems such as chalcopyrite band gap prediction,[89] band gap engineering,[90] scintillator discovery,[91] mechanical properties of alloys[92] and phosphor data mining.[93]

An example of the application of such techniques is the use of principal component linear regression analysis by Curtarolo et al. to predict energies of compounds.[94] In this study, the energy of a compound was demonstrated to be predictable by linear regression from the energies of other compounds in the same chemical system. Another example pertains to new piezoelectrics discovery: Balachandran et al. reduced a set of 30 candidate descriptors down to 6 using principal components analysis.[95] This reduced set of descriptors was subsequently incorporated into a linear model that predicted the Curie temperature at the morphotropic phase boundary, including two new possible piezoelectric materials (see Fig. 6).

E. Kernel ridge regression (KRR)

Linear techniques might not adequately capture complex relationships in the data. One method to capture nonlinear relationships is to retain linear models, but to transform the descriptors in nonlinear ways (e.g., by taking the square or logarithm of descriptors) prior to fitting the linear model. This idea forms the essence of KRR, which leverages two principles beyond standard

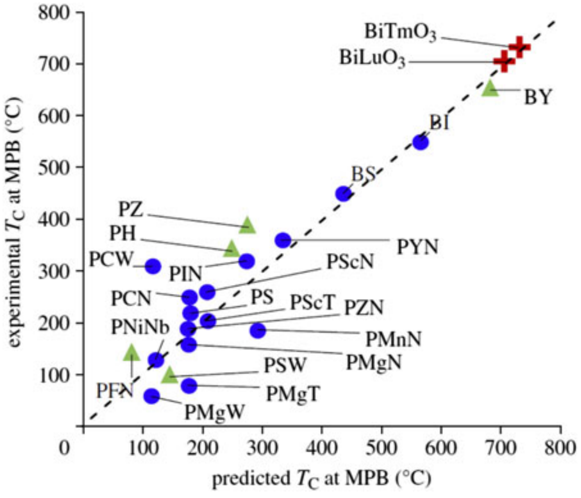

FIG. 6. Predicted versus experimental Curie temperature for a series of piezoelectric materials. Blue points are the training set, green triangle the test set and red cross predicted new piezoelectrics. Image from Ref. 95. ©The Royal Society, reprinted with permission.

linear regression. First, KRR uses the ridge regression principle of *regularizing* the fitted coefficients by adding a penalty term that scales with their l_2 norm. The regularization term, used to control over-fitting, is particularly important for kernel-based methods like KRR because of the higher number of dimensions being used for the fit. The second principle of KRR is to use a *kernel* to lift the input descriptors into a higher dimensional space. The mathematics of kernels can be found elsewhere,[96] but the general principle is that problems that can be posed as inner products can be mapped to higher dimensions through kernels that transform the inputs. This allows KRR to efficiently minimize the ridge regression error function even when the number of dimensions in the transformed feature vectors is very high and perhaps even in excess of the number of observations. There are many types of kernels that can be used for this higher-dimensional mapping, including polynomial, gaussian, exponential, and others. The choice of kernel function and the magnitude of the regularization parameter will affect the fitting and must often be tuned, e.g., through cross-validation.

The KRR model is *nonparametric*, i.e., the prediction for a new feature vector is not expressed through a typical linear equation expressed in terms of transformed features but rather by projecting that targeted feature vector onto the solution space through dot products with the training data. KRR has been used to model the properties of small molecules (atomization energies[4] and multiple properties of 1D chains[97]) and have recently been applied to solids. For example, Shütt et al. used KRR to predict the DOS at the Fermi energy using a sample of 7000 compounds, finding a high correlation (error of roughly 6% of the range of values) that depends on the types of electronic orbitals involved.[98] A second example is from Faber et al., who used KRR to predict formation energies using a sample of almost 4000 compounds, achieving errors as low as 0.37 eV/atom.[71] In both instances, a critical part of the analysis was deciding how to represent crystalline materials through descriptors; Shütt et al. leveraged a form similar to radial distribution functions,[98] whereas Faber et al. modified a Coulomb matrix form that was developed for molecules.[71]

F. Nonlinear techniques based on trees

There exist several methods designed to directly capture nonlinear functions of descriptors that produce an output. One such method is that of neural networks, in which descriptor values serve as "input nodes" that form a weighted graph through "hidden nodes" and end in a set of "output" nodes, with the weightings describing (in a nonlinear way) how the inputs relate to the outputs and are trained by the data. Neural nets have been applied in materials science[99–102] and are experiencing a general resurgence in the form of "deep learning"; however, they suffer from a lack of interpretability, and large data sets are usually needed to adequately train them.

An alternative nonlinear technique is that of tree-based models. Tree-based methods iteratively split variables into successively smaller groups, typically in a way that maximizes the homogeneity within each branch. In this way, trees are similar to clustering, but they are *supervised* (split on the value of a known "response" variable but using decisions only on the "predictors") whereas clustering is *unsupervised* (finds internal relationships between all variables, without distinction between "predictor" and "response"). A tree model makes a series of decisions based on descriptor values that successively move down the branches, until one reaches a "leaf" (ending) node within which the values of the responses are sufficiently homogeneous to make a prediction.

One field in which tree-based methods have been applied is thermoelectrics design. Carrete et al. have constructed decision trees[103] that separate half-heusler structures into high and low probability of possessing high figure-of-merit based on decisions regarding electronegativity difference, atomic number difference, and periodic table column (see Fig. 7). This study also used the tree-based method of random forests (described next) to predict thermal conductivity.

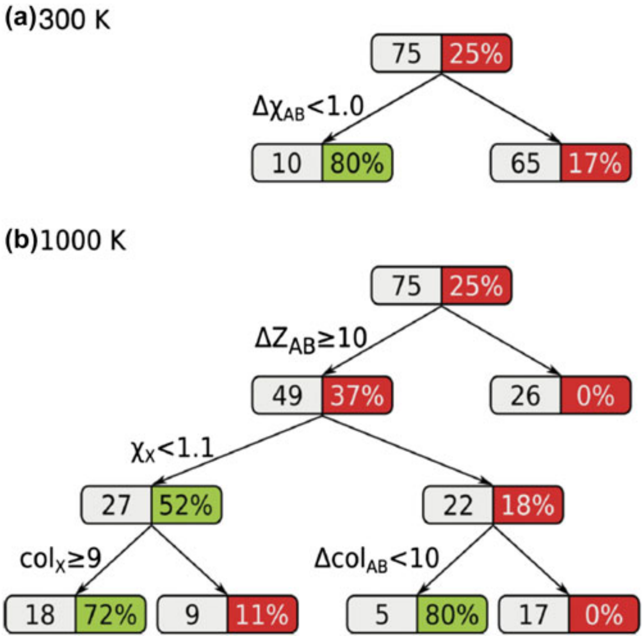

FIG. 7. Results from a decision tree algorithm on a data set of 75 half heusler compounds; labels above the arrows represent decisions, and nodes are divided into number of compounds and fraction of compounds remaining that possess high ZT (the thermoelectric figure-of-merit). The decision tree highlights the most important factors leading to high ZT materials for two different operating temperatures (300 and 1000 K). Image from Ref. 103. ©Wiley-VCH Verlag GmbH & Co.,reprinted with permission.

There exist many techniques for improving the predictive performance of trees, such as *pruning* (removing endpoint branches) to avoid over-fitting and *smoothing* to reduce variance in prediction (so that small changes in input data do not result in abrupt changes to the output). In recent years, *ensemble-based* techniques based on averaging the predictions of several models have gained much popularity. For example, the technique of *random forests*[104] fits many individual tree models (as the name suggests) and then averages or takes a vote on the results. To prevent each tree in the random forest from producing the same result, each model is (i) fit using only a subset of the training data and (ii) employs only a subset of the possible descriptors. We note that when only (i) is used, the technique is termed *bagging* and falls under the general technique of *bootstrapping*. The random forest method results in a mix of slightly different "opinions" that are used to form a coherent consensus that reduce bias of a single model and produces smoother results than a single tree. We note that ensembling techniques are more general than tree-based methods and can indeed mix many different types of models. For example, Meredig et al.[105] combined a heuristic approach (based on relating ternary formation energies to a weighted sum of binary formation energies) with a regression approach based on ensembles of decision trees to build a model linking composition to formation energy.

Ensembles are effective because a majority vote of many models can produce better results than any individual model, provided that the errors in the individual models are not highly correlated. For example, by combining 5 different models that are each only 75% accurate in predicting a binary outcome (with random errors) into a majority vote ensemble, the overall probability of a correct classification increases to approximately 90%. Unfortunately, such models can be difficult to interpret and might be overfit if a proper machine learning design (training, validation, test sets)[106] is not followed. Nevertheless, the high level of performance that can be achieved will likely make ensembles popular in future materials science studies.

G. Recommendation engines

Another class of data mining techniques is based on finding associations in data and can be used to create a ranked list of suggestions or probable outcomes. Using these techniques, one can derive useful associations in materials properties in the same way that a retailer can predict that a shopper who purchases cereal is likely to also purchase milk.

Indeed, some of the first applications of advanced machine learning algorithms in materials science were based on this idea. Fischer et al. developed a Bayesian statistics-based algorithm trained on a large crystallographic experimental database (the Pauling file[8]) to build correlations between existing binary crystal structures and proposed ones.[107] The model used the associations to build predictive models, such as: if A_1B_1 composition of two elements is known to form the rock salt structure, then what is the probability that a new proposed A_2B_1 compound of those elements crystallizes in the Mg_2Si prototype structure? Fischer et al. analyzed such associations to derive probabilistic statements on what structures are likely to form at unexplored compositions based on what structures are observed at measured compositions. This approach was later extended to ternary systems and in particular ternary oxides to perform an automatic search for new compounds.[65] Among the reported successes is the experimental confirmation of the formation of $SnTiO_3$ in an ilmenite structure rather than perovskite (as was previously proposed in the literature[108]).

Subsequently, another compound prediction data mining-based algorithm, specifically designed for data-sparse regions such as quaternary compounds, was proposed by Hautier et al.[66] This method is based on the simple assumption that if a compound forms in a given crystal structure (e.g., a perovskite), replacing one of its ions by an ion that is chemically similar (e.g., in the same column of the periodic table or in the same size) is likely to lead to a new stable compound in the same structure. The idea was implemented within a rigorous mathematical framework and trained on compounds present in the ICSD database.[6] Fig. 8 represents a map of pair correlations between ions obtained from the probabilistic model. The map indicates which ions are likely (red) and unlikely (blue) to substitute for one another. Some compounds predicted by this work (mainly rooted in an interest in the Li-ion battery field) and confirmed experimentally include: $Li_9V_3(P_2O_7)_3PO_4)_2$,[76,109,110] $A_3M(CO_3)(PO_4)$ (A = alkali, M = TM),[111,112] $LiCoPO_4$,[113] Li_3CuPO_4,[114] and $LiCr_4(PO_4)_3$.[115] An online version of this tool is available as the Materials Project "Structure Predictor" app. More recent work by Yang et al. has further built on this concept and proposed a composition similarity mapping.[116]

Another example of associative learning is in the prediction of photocatalyst materials. Castelli et al. computed several key photocatalytic properties using DFT for over 18,000 ABO_3 oxides, then used association techniques to derive the probability of an arbitrary *oxynitride* (i.e., ABO_2N) to meet the same photocatalytic criteria.[83] By learning the A and B element identities, most associated with good photocatalytic behavior in the oxides data and transferring the same assumptions for oxynitrides, they were able to find 88% of the target oxynitride systems by searching only 1% of the data set.

Such analyses are culminating in more general "recommender" systems that can suggest new compounds

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

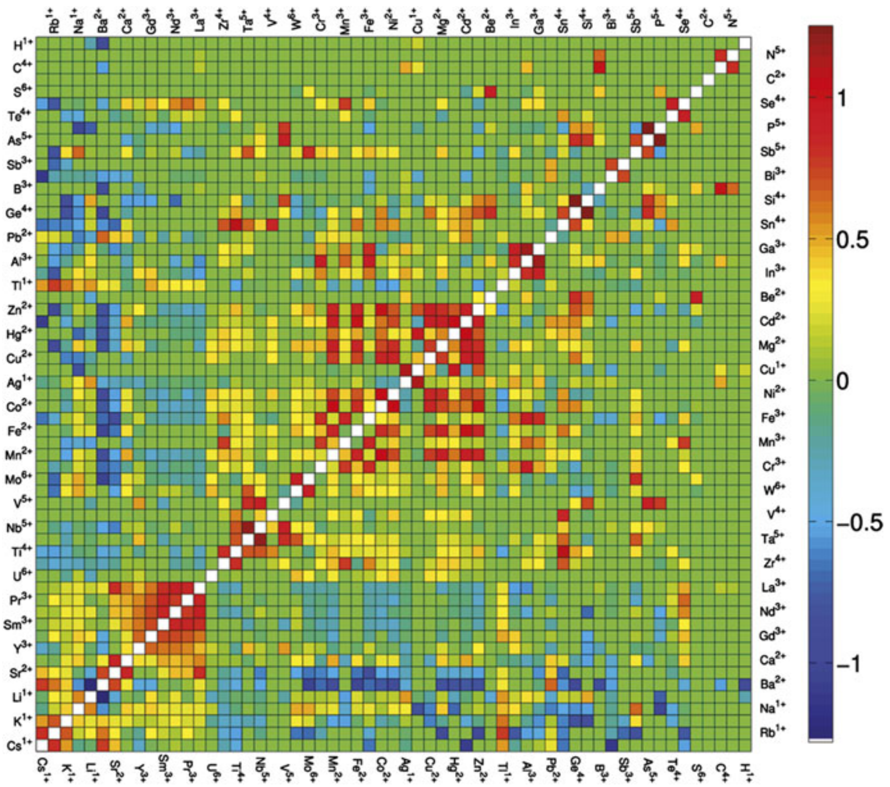

FIG. 8. Map of the pair correlation for two ions to substitute within the ICSD database. Positive values indicate a tendency to substitute, whereas negative values indicate a tendency to not substitute. The symmetry of the pair correlation ($g_{ab} = g_{ba}$) is reflected in the symmetry of the matrix. Image from Ref. 66. ©The American Chemical Society, reprinted with permission.

based on observed data. For example, Citrine Informatics has built a recommendation system for suggesting new and unconventional thermoelectric materials.[117] Another example is that of Seko et al.,[73] who used the method of kriging, based on Gaussian processes, to search for systems with high melting point based on observed data. The same technique was recently applied to identify low thermal conductivity materials.[118] A particularly fruitful field of study in the future may be in the combination of such recommender systems with high-throughput computation. In such a scheme, the results of an initial set of computations would produce the initial data set needed for a recommender system to make further suggestions. These new suggestions would be computed automatically and then fed back into the recommender, thereby creating a closed loop that continually produces interesting candidates for follow up.

IV. APPLICATION EXAMPLE: VALENCE AND CONDUCTION BAND CHARACTER

The combination of open materials databases, programmatic APIs, and data mining techniques has recently exposed new and unprecedented opportunities to apply materials informatics. We next demonstrate work that analyzes the character of valence and conduction bands in materials over a large data set. This study was performed using only free software tools: the Materials Project database,[38] the MAPI programmatic API,[59] the pymatgen code,[53] and the Python and R programming languages (including the BradleyTerry2 package[119] implemented in R).

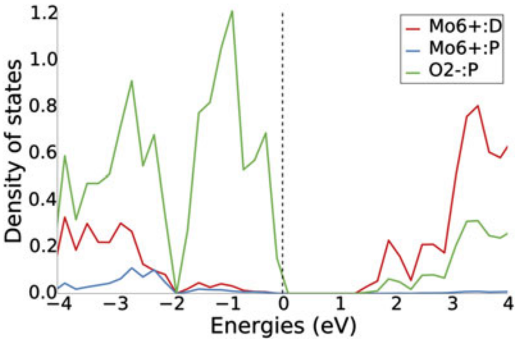

FIG. 9. Example of projected DOS used in the study (MoO_3, entry mp-18856 in the Materials Project[38]). The states near the valence band are dominated by O^{2-}:*p*, whereas those near the conduction band are dominated by Mo^{6+}:*d*. Over 2500 such materials are used to assess statistics on valence and conduction band character in this study.

The aim of our study is to develop a probabilistic model for predicting the dominant *character* (element, charge state, and orbital type) of the electronic states near the valence band maximum (VBM) and conduction band minimum (CBM) over a broad range of compounds. Many materials properties, including the band gap (important for light capture) and Seebeck coefficient (important for thermoelectrics) depend critically on the details of these band edges. Understanding the character of these states allows one to determine the type of modifications needed to achieve desired properties.

Our strategy for developing a probabilistic model of VBM and CBM character can be outlined as follows:

(i) Select a computational data set from the Materials Project database and download the projected DOS information using the MAPI. The projected DOS contains the relative contributions of all the elements and orbitals in the material to the DOS at each energy level (see Fig. 9).

(ii) For each material, assign valence states to each element using a bond valence method; remove materials with mixed or undetermined valence.

(iii) Assess the contribution of each combination of element, valence, and orbital (e.g., O^{2-}:p, hereafter called an **ionic orbital**) to the VBM and CBM, thereby assigning a *score* to each ionic orbital (based on Fermi weighting, see Supplementary Info) toward the VBM and CBM in that material.

(iv) For each material, and separately for the VBM and CBM, determine the higher score amongst all pairs of ionic orbitals. This can be considered a set of "competitions" within that material for greater contribution to the VBM and CBM for which there are binary win/loss records between pairs of ionic orbitals (e.g., O^{2-}:*p* contributes more to the VBM than F^{1-}:*p*). When iterated over all materials in the data set, this will result in a comprehensive set of pairwise rankings regarding which ionic oribtals tend to dominate the VBM/CBM.

(v) Use a Bradley–Terry model[120] to transform pairwise assessments of ionic orbitals into a single probabilistic ranking of which ionic orbitals are most likely to form the VBM and CBM states. This technique is useful because each material provides only a limited view (i.e., comparisons between the few elements contained in that material) of the overall ranking between all ionic orbitals. For example, this technique is one way to determine an absolute ranking of sports teams based only on pairwise competition data.

The details of our methodology, including considerations such as the treatment of *d* orbitals in the Materials Project dataset, are presented in the Supplementary Information.

The results of our study for a selected set of ionic orbitals are presented in Fig. 10; the full results are presented in the Supplementary Information. Each data point in Fig. 10 represents the probability that the ionic orbital listed on the *y* axis will have a greater contribution to the specified band edge than the orbital listed on the *x* axis. The ionic orbitals that tend to dominate are ordered top-to-bottom along the *y* axis, and left-to-right along the *x* axis. For example, the data indicate that in a material containing both Cu^{1+}:*d* and Fe^{3+}:*d* states, the VBM is highly likely to be dominated by Cu^{1+}:*d* (probability close to 1.0) where as the CBM is more likely to be dominated by Fe^{3+}:*d*. We note that our model should not be interpreted as a ranking of energy levels in compounds. For example, an ionic orbital may have a low VBM score either because the energy tends to be too high (it forms a conduction band) or too low (too deep in the valence band).

As an example of how this type of analysis can be used in materials science problems, the absolute VBM position of a material has been demonstrated to be correlated with its ease to be p-type.[121–123] Oxides with purely oxygen valence band character tend to have a too low valence band to be p-type dopable. The design principle of using cations that will hybridize with oxygen to lift the valence band up and facilitate p-type doping is an important paradigm in the development of new p-type oxides and especially transparent conducting oxides (TCOs). The presence in our data set of Cu^{1+} or Ag^{1+} as cations that compete strongly with oxygen is consistent with the use of these elements in p-type TCOs.[124–129] Notably, other 3*d* ions in our analysis, including V^{3+} and $Mn^{2+,}$ are also known to form p-type oxides.[130] This analysis can therefore be used as a first step toward the more systematic listing of ions necessary to form p-type oxides.

In some cases, the data do not match our intuition; for example, we expect that S^{2-}:*p* would be ranked higher in the valence band than O^{2-}:*p*. Counter to our result that O^{2-}:*p* and S^{2-}:*p* are ranked similarly, in all specific examples in our data set for which a single material simultaneously contained both O^{2-}:*p* and S^{2-}:*p* orbitals, the greater VBM contribution came from S^{2-}:*p*. This discrepancy suggests that a consistent universal ranking of ionic orbitals may not exist. For example, when considering compounds containing Mn^{2+}:*d*, the VBM is composed mostly of S^{2-}:*p* in only 20% of cases versus 88% for O^{2-}:*p* (see the Supplementary Info). An example of the latter is $MnPO_4$ with spacegroup *Pmnb* (mp-777460 on the Materials Project[38] web site), in which O^{2-}:*p* dominates the VBM, whereas Mn^{2+}:*d* forms the CBM. Thus, interactions between ionic orbitals can be in conflict depending on a specific material's physics. Possible improvements to the model may include more heavily weighting direct competitions, or taking into account the relative energies between ionic orbitals. Nevertheless, overall our scores are in good agreement with known principles and provide general guidelines for engineering band edges.

Our model for predicting VBM and CBM characters is one example of how it is now possible to leverage open materials data sets and software tools to provide rapid and

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

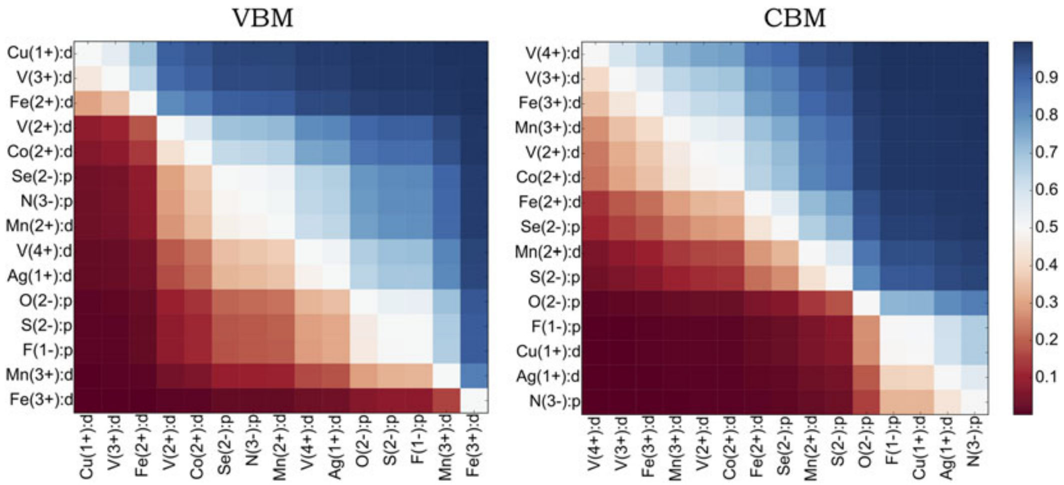

FIG. 10. Pairwise probability for the ionic orbital listed to the left to have a greater contribution to a band edge then the ionic orbital listed at the bottom. The VBM edge data are listed to the left, and the CBM data to the right. Only a selected set of ionic orbitals are depicted; the full data set is in the Supplementary Information.

data-driven models for problems normally guided by intuition and experience. However, it should be noted that the results of such approaches are still limited by the accuracy of the underlying data set. In the case of this analysis, the generalized gradient approximation with +U correction (GGA+U) method was used by the Materials Project to handle localized/correlated *d* states, but this method is not universally accurate. Nevertheless, the fact that today, these models can be built on top of already existing data sets and software libraries speaks to the potential democratic power of materials informatics.

V. DISCUSSION AND CONCLUSION

Data mining and informatics-based approaches present new opportunities for materials design and understanding. As the amount of publicly available materials data grows, such techniques will be able to extract from these data sets scientific principles and design rules that could not be determined through conventional analyses. In this paper, we reviewed some of these emerging materials databases along with several techniques and examples of how materials informatics can contribute to materials discovery. We also demonstrated how one can already combine open data and software tools to produce an analysis predicting the character of a compound's valence and conduction bands. However, despite these opportunities, several challenges remain that have limited the impact of materials informatics approaches thus far and will no doubt be the subject of future work.

The first remaining challenge is for the community to gain experience in navigating the challenges of using large data resources. Even as the number of databases grows and as programmatic APIs to the data become available, the fraction of users that extract and work with large data sets remains relatively small. Furthermore, when the desired data are not available from a single source, it is difficult to query over multiple resources and to combine information from different databases. It is especially challenging to match computations against experimental measurements because data regarding the crystal structure and other conditions under which the experiments were performed is usually missing. Working with computational data alone is also tricky; researchers must understand the error bars of the simulation method, which can sometimes be quite high and difficult even for expert users to estimate. Thus, even though the situation for materials data is improving rapidly, today one must still be somewhat of an expert user to exploit these resources.

The second challenge is the development of materials descriptors for crystalline, periodic solids. While there has been progress in the last several years, this area is still ripe for new ideas. Today, we still do not possess automatic algorithms that can describe crystals using descriptors that would typically be used by a domain expert. Such descriptors could include the nature and connectivity of local environments (e.g., “edge-sharing tetrahedra” versus “corner-sharing octahedra”), qualitative assessments of structure (e.g., “close-packed” versus “layered” versus “1D channels”), or crystal prototype (e.g., “double perovskite” or “ordering of rock salt lattice”). Crystal structure data are quite complex and varied. Without such algorithms, it is difficult for researchers to describe a crystal to a machine learning algorithm using constructs proven to be successful by decades of materials science.

Finally, the third challenge is in assessing the appropriateness and transferability of machine learning models. Typically, such assessments are made using a performance-oriented metric such as cross-validation error. However, there are many reasons why such performance metrics might be misleading. The first issue is that the cross-validation error can be affected by the type of cross-validation (e.g., leave-one-out versus *n*-fold)[131] as well as the design of how models are selected and how the data are split for fitting (i.e., training, validation, and test sets).[106] However, perhaps an even bigger concern is in controlling and reporting *sampling errors*, i.e., in training and cross-validating a model on a sample that is not representative of the full population. Examples of potential sampling error include (i) building and validating a model on a data set of binary compounds and subsequently applying that error estimation to ternary and quaternary compounds, (ii) building and validating a model's performance for a limited number of highly symmetric structure types, and then applying the same model to predict the behavior of dissimilar crystal structures, and (iii) training the model only on thermodynamically stable compounds and then applying that model to unstable compounds. In many cases, it is very difficult to obtain data for samples that are truly representative of the prediction space, and better methods to estimate error and applicability are needed for these situations.

This third issue of knowing how much weight to put in a machine learning model is particularly important because the models that achieve the lowest cross-validation error are often the most complex (e.g., neural nets or random forests) and can be impossible to interpret scientifically. Should materials scientists trust a prediction from an impenetrable machine learning model when it is in conflict with the intuition afforded by a more classical, interpretable model? Part of the distinction lies in whether the goal of machine learning models is to simply be *predictive* (capable of making useful forecasts) or whether additional weight should be afforded to models that are causal (confirming that a factor is truly the reason behind an effect) or mechanistic (reproduce the physics behind an effect). Ideally, machine learning models will not only try to answer specific questions accurately but will also prove useful in leading us to better questions and to new types of analyses (as was the case for the periodic table).

While it will take some time to truly develop solutions to all of these challenges, informatics is already making headway into materials science, and data-driven methods will no doubt form a major area in the study of materials in the future.

ACKNOWLEDGMENTS

This work was intellectually led by the Materials Project (DOE Basic Energy Sciences Grant No. EDC-BEE). Work at the Lawrence Berkeley National Laboratory was supported by the U.S. Department of Energy Office of Science, Office of Basic Energy Sciences Department under Contract No. DE-AC02-05CH11231. GH acknowledges financial support from the European Union Marie Curie Career Integration (CIG) grant HT4TCOs PCIG11-GA-2012-321988. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility.

REFERENCES

1. G. Hautier, A. Jain, and S.P. Ong: From the computer to the laboratory: Materials discovery and design using first-principles calculations. *J. Mater. Sci.* **47**(21), 7317–7340 (2012).
2. K. Rajan and P. Mendez: Materials informatics. *Mater. Today* **8**(10), 38–45 (2005).
3. M. Rupp, E. Proschak, and G. Schneider: Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.* **47**(6), 2280–2286 (2007).
4. M. Rupp, A. Tkatchenko, K.-R. Müller, V. Lilienfeld, and O. Anatole: Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
5. K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. Von Lilienfeld, A. Tkatchenko, and K.R. Müller: Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
6. G. Bergerhoff, R. Hundt, R. Sievers, and I.D. Brown: The inorganic crystal-structure database. *J. Chem. Inf. Comput. Sci.* **23**(2), 66–69 (1983).
7. F.H. Allen: The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **58**, 380–388 (2002).
8. P. Villars: The linus pauling file (LPF) and its application to materials design. *J. Alloys Compd.* **279**(1), 1–7 (1998).
9. R.D. Shannon: Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr., Sect. A: Found. Adv.* **32**(5), 751–767 (1976).
10. I.D. Brown and D. Altermatt: Bond-valence parameters obtained from a systematic analysis of the inorganic crystal structure database. *Acta Crystallogr., Sect. B: Struct. Sci.* **244**(2), 244–247 (1985).
11. M. O'Keefe and N.E. Brese: Bond–valence parameters for solids. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **47**, 192–197 (1991).
12. I. Brown and K. Wu: Empirical parameters for calculating cation-oxygen bond valences. *Acta Crystallogr., Sect. B: Struct. Sci.* **32**(31563), 1957–1959 (1976).
13. I.D. Brown: On the geometry of OH. . .O hydrogen bonds. *Acta Crystallogr., Sect. A: Found. Adv.* **32**(31563), 24–31 (1976).
14. D. Yu and D. Xue: Bond analyses of borates from the inorganic crystal structure database. *Acta Crystallogr., Sect. B: Struct. Sci.* **62**, 702–709 (2006).
15. A.L. Mackay: The statistics of the distribution of crystalline substances among the space groups. *Acta Crystallogr.* **22**, 329–330 (1967).
16. V.S. Urusov and T.N. Nadezhina: Frequency distribution and selection of space groups in inorganic crystal chemistry. *J. Struct. Chem.* **50**, 22–37 (2009).
17. S.C. Abrahams: Inorganic structures in space group P3m1; Coordinate analysis and systematic prediction of new ferroelectrics. *Acta Crystallogr., Sect. B: Struct. Sci.* **64**, 426–437 (2008).

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

18. M. Avdeev, M. Sale, S. Adams, and R.P. Rao: Screening of the alkali-metal ion containing materials from the inorganic crystal structure database (ICSD) for high ionic conductivity pathways using the bond valence method. *Solid State Ionics* 2–5 (2012).
19. O. Muller and R. Roy: *The Major Ternary Structural Families* (Springer-Verlag, New York, 1974).
20. D.G. Pettifor: The structures of binary compound: I. Phenomenological structure maps. *J. Phys. C: Solid State Phys.* **19**, 285–313 (1986).
21. D.G. Pettifor: Structure maps in alloy design. *J. Chem. Soc., Faraday Trans.* **86**(8), 1209–1213 (1990).
22. D.G. Pettifor: Structure maps revisited. *J. Phys.: Condens. Matter* **15**, 13–16 (2003).
23. D. Morgan, J. Rodgers, and G. Ceder: Automatic construction, implementation and assessment of Pettifor maps. *J. Phys.: Condens. Matter* **15**, 4361–4369 (2003).
24. C.S. Kong, W. Luo, S. Arapan, P. Villars, S. Iwata, R. Ahuja, and K. Rajan: Information-theoretic approach for the discovery of design rules for crystal chemistry. *J. Chem. Inf. Model* **52**, 1812–1820 (2012).
25. P.S. White, J.R. Rodgers, and Y. Le Page: Crystmet: A database of the structures and powder patterns of metals and intermetallics. *Acta Crystallogr., Sect. B: Struct. Sci.* **58**, 343–348 (2002).
26. P. Villars and K. Cenzual: Pearsons crystal data: Crystal structure database for inorganic compounds (ASM International/Material Phases Data System, Vitznau, Switzerland, 2010).
27. L. Glasser: Crystallographic information resources. *J. Chem. Educ.* (2015). acs.jchemed.5b00253.
28. SpringerMaterials: The Landolt-Börnstein database. www.springermaterials.com/.
29. C. Bale, E. Bélisle, P. Chartrand, S. Decterov, G. Eriksson, K. Hack, I-H. Jung, Y-B. Kang, J. Melançon, A. Pelton, C. Robelin, and S. Petersen: FactSage thermochemical software and databases recent developments. *Calphad* **33**(2), 295–311 (2009).
30. P. Linstrom and W. Mallard: *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (National Institute of Standards and Technology, Gaithersburg MD 20899, 2015).
31. L. MatWeb: MatWeb, Material property data, Data base of materials data sheets.
32. MatNavi: NIMS materials database. http://mits.nims.go.jp/index_en.html. (2014).
33. O. Kubaschewski, C.B. Alcock, and P.J. Spencer: Thermochemical Data, in: *Materials Thermochemistry*, 6th ed. (Pergamon Press, Oxford, 1993); ch. 5, p. 376.
34. H. Okamoto: In *Handbook of Ternary Alloy Phase Diagrams*, P. Villars, A. Prince, and H. Okamoto eds.; (ASM International: OH, 1995); pp. 10378–10379.
35. P. Hohenberg and W. Kohn: Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
36. W. Kohn and L. Sham: Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, 1133–1138 (1965).
37. M.D. Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C.K. Ande, S.V.D. Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, and M. Asta: Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 1–13 (2015).
38. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson: Performance of genetic algorithms in search for water splitting perovskites. *APL Mater.* **1**, 011002 (2013).
39. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy: Aflowlib.org: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
40. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton: Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**(11), 1501–1509 (2013).
41. J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sanchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, and A. Aspuru-Guzik: The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**(17), 2241–2251 (2011).
42. C. Ortiz, O. Eriksson, and M. Klintenberg: Data mining and accelerated electronic structure theory as a tool in the search for new functional materials. *Comput. Mater. Sci.* **44**(4), 1042–1049 (2009).
43. E. Blokhin, L. Pardini, F. Mohamed, K. Hannewald, L. Ghiringhelli, P. Pavone, C. Carbogno, J-C. Freytag, C. Draxl, and M. Scheffler: The NoMaD Repository. http://nomad-repository.eu/.
44. V. Stevanović, S. Lany, X. Zhang, and A. Zunger: Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. *Phys. Rev. B: Condens. Matter Mater. Phys.* **85**(11), 1–12 (2012).
45. D.D. Landis, J.S. Hummelshøj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. Norskov, and K. Jacobsen: The computational materials repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).
46. J.S. Hummelshøj, F. Abild-Pedersen, F. Studt, T. Bligaard, and J.K. Nørskov: CatApp: A web application for surface chemistry and heterogeneous catalysis. *Angew. Chem., Int. Ed. Engl.* **51**(1), 272–274 (2012).
47. A. Togo and I. Tanaka: First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
48. A. Togo: PhononDB at Kyoto University (http://phonondb.mtl.kyoto-u.ac.jp).
49. P. Gorai, D. Gao, B. Ortiz, S. Miller, S.A. Barnett, T. Mason, Q. Lv, V. Stevanović, and E.S. Toberer: Te design lab: A virtual laboratory for thermoelectric material design. *Comput. Mater. Sci.* **112**, 368–376 (2016).
50. G. Yuan and F. Gygi: Estest: A framework for the validation and verification of electronic structure codes. *Comput. Sci. Discovery* **3**(1), 015004 (2010).
51. H.E. Pence and A. Williams: ChemSpider: An online chemical information resource. *J. Chem. Educ.* **87**(11), 1123–1124 (2010).
52. L. Lin: Materials databases infrastructure constructed by first principles calculations: A review. *Mater. Perform. Charact.* **4**, MPC20150014 (2015).
53. S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, and G. Ceder: Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
54. S. Bahn and K. Jacobsen: An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **4**(3), 56–66 (2002).
55. S. Curtarolo, W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, and D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery, *Comput. Mater. Sci.* **58**, 218–226 (2012).
56. G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky: AiiDA: automated interactive infrastructure and database for

computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).

57. A. Jain, S. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G-M. Rignanese, G. Hautier, D. Gunter, and K. Persson: FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput. Pract. Exp.* **27**, 5037–5059 (2015).
58. R.T. Fielding: Architectural styles and the design of network-based software architectures. Ph.D. Dissertation, University of California, Irvine, 2000.
59. S.P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K.A. Persson: The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
60. T. Hastie, R. Tibshirani, and J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics*, 2nd ed. (Springer, New York, 2009); ch. 4, pp. 80–113.
61. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, and M. Scheffler: Big data of materials science : Critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
62. K. Yang, W. Setyawan, S. Wang, M. Buongiorno Nardelli, and S. Curtarolo: A search model for topological insulators with high-throughput robustness descriptors. *Nat. Mater.* **11**(7), 614–619 (2012).
63. H. Burzlaff and H. Zimmermann: On symmetry classes of crystal structures. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **65**, 456–465 (2009).
64. R. Allmann and R. Hinek: The introduction of structure types into the inorganic crystal structure database ICSD. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **63**, 412–417 (2007).
65. G. Hautier, C.C. Fischer, A. Jain, T. Mueller, and G. Ceder: Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**(12), 3762–3767 (2010).
66. G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder: Data mined ionic substitutions for the discovery of new compounds. *Inorg. Chem.* **50**(17), 656–663 (2010).
67. J. Behler and M. Parrinello: Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**(14), 146401 (2007).
68. L. Yang, S. Dacek, and G. Ceder: Proposed definition of crystal substructure and substructural similarity. *Phys. Rev. B: Condens. Matter Mater. Phys.* **90**(5), 054102 (2014).
69. A.R. Oganov and M. Valle: How to quantify energy landscapes of solids. *J. Chem. Phys.* **130**(10), 104504 (2009).
70. O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo: Materials cartography: Representing and mining material space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2014).
71. F. Faber, A. Lindmaa, O.A. von Lilienfeld, and R. Armiento: Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1–8 (2015).
72. C.S. Kong, S.R. Broderick, T.E. Jones, C. Loyola, M.E. Eberhart, and K. Rajan: Mining for elastic constants of intermetallics from the charge density landscape. *Phys. B* **458**, 1–7 (2015).
73. A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka: Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids. *Phys. Rev. B: Condens. Matter Mater. Phys.* **89**, 054303 (2014).
74. M. Schmidt and H. Lipson: Distilling free-form natural laws from experimental data. *Science* **324**(5923), 81–85 (2009).
75. A. Jain, G. Hautier, C.J. Moore, S.P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, and S. Ping Ong: A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295–2310 (2011).
76. G. Hautier, A. Jain, S.P. Ong, B. Kang, C. Moore, R. Doe, and G. Ceder: Phosphates as lithium-ion battery Cathodes: An evaluation based on high-throughput ab initio calculations. *Chem. Mater.* **23**, 3508–3945 (2011).
77. A. Jain, G. Hautier, S.P. Ong, S. Dacek, and G. Ceder: Relating voltage and thermal safety in Li-ion battery cathodes: a high-throughput computational study. *Phys. Chem. Chem. Phys.* **17**, 5942–5953 (2015).
78. S.P. Ong, A. Jain, G. Hautier, B. Kang, and G. Ceder: Thermal stabilities of delithiated olivine MPO_4 (M = Fe, Mn) cathodes investigated using first principles calculations. *Electrochem. Commun.* **12**(3), 427–430 (2010).
79. N.A. Godshall, I.D. Raistrick, and R.A. Huggins: Relationships among electrochemical, thermodynamic, and oxygen potential quantities in lithium-transition metal-oxygen molten salt cells. *J. Electrochem. Soc.* **131**(3), 543 (1984).
80. R. Xu and D. Wunsch II: Survey of clustering algorithms, neural networks, *IEEE Trans. Neural Networks* **16**, 645–678 (2005).
81. G. Gan, C. Ma, and J. Wu: *Data clustering: theory, algorithms, and applications*, Vol. **20** (Society for Industrial and Applied Mathematics, Philadelphia, 2007).
82. B. Meredig and C. Wolverton: Dissolving the periodic table in cubic zirconia: Data mining to discover chemical trends. *Chem. Mater.* **26**(6), 1985–1991 (2014).
83. I.E. Castelli and K.W. Jacobsen: Designing rules and probabilistic weighting for fast materials discovery in the perovskite structure. *Modell. Simul. Mater. Sci. Eng.* **22**(5), 055007 (2014).
84. S.R. Broderick, H. Aourag, and K. Rajan: Classification of oxide compounds through data-mining density of states spectra. *J. Am. Ceram. Soc.* **94**(9), 2974–2980 (2011).
85. R. Andersen: *Modern Methods for Robust Regression* (Sage, Los Angeles, 2008).
86. J.R. Chelikowsky and K.E. Anderson: Melting point trends in intermetallic alloys. *J. Phys. Chem. Solids* **48**(2), 197–205 (1987).
87. R. Tibshirani: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
88. H. Zou and T. Hastie: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**(2), 301–320 (2005).
89. P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan: Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
90. S. Srinivasan and K. Rajan: "Property phase diagrams" for compound semiconductors through data mining. *Materials* **6**(1), 279–290 (2013).
91. C.S. Kong and K. Rajan: Rational design of binary halide scintillators via data mining. *Nucl. Instrum. Methods Phys. Res., Sect. A* **680**, 145–154 (2012).
92. I. Toda-Caraballo, E.I. Galindo-Nava, and P.E.J. Rivera-Díaz-Del-Castillo: Unravelling the materials genome: Symmetry relationships in alloy properties. *J. Alloys Compd.* **566**, 217–228 (2013).
93. W.B. Park, S.P. Singh, M. Kim, and K-S. Sohn: Phosphor informatics based on confirmatory factor analysis. *ACS Comb. Sci.* 150408124118005 (2015).
94. S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder: Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91**(13), 135503 (2003).
95. P.V. Balachandran, S.R. Broderick, and K. Rajan: Identifying the 'inorganic gene' for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A* **467**, 2271–2290 (2011).

https://doi.org/10.1557/jmr.2016.80 Published online by Cambridge University Press

96. N. Cristianini and J. Shawe-Taylor: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press, 2000).
97. G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad: Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
98. K.T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.R. Müller, and E.K.U. Gross: How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **89**, 1–5 (2014).
99. R. Jalem, M. Nakayama, and T. Kasuga: An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. *J. Mater. Chem. A* **2**(3), 720 (2014).
100. F. Pettersson, C. Suh, H. Saxen, K. Rajan, and N. Chakraborti: Analyzing sparse data for nitride spinels using data mining, neural networks, and multiobjective genetic algorithms. *Mater. Manuf. Processes* **24**(1), 2–9 (2009).
101. D. Scott, S. Manos, and P. Coveney: Design of electroceramic materials using artificial neural networks and multiobjective evolutionary algorithms. *J. Chem. Inf. Model.* **48**, 262–273 (2008).
102. Y. Zhang, S. Yang, and J. Evans: Revisiting Hume-Rotherys rules with artificial neural networks. *Acta Mater.* **56**(5), 1094–1105 (2008).
103. J. Carrete, N. Mingo, S. Wang, and S. Curtarolo: Nanograined half-heusler semiconductors as advanced Thermoelectrics: An ab initio high-throughput statistical study. *Adv. Funct. Mater.* **24**, 7427–7432 (2014).
104. A. Liaw and M. Wiener: Classification and regression by randomForest. *R News* **2**(3), 18–22 (2002).
105. B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton: Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **89**(9), 094104 (2014).
106. R. Bell, Y. Koren, and C. Volinsky: Chasing $1,000,000: How We Won The Netflix Progress Prize, *Statistical Computing and Statistical Graphics Newsletter* **18**(2), 4–12 (2007).
107. C.C. Fischer, K.J. Tibbetts, D. Morgan, and G. Ceder: Predicting crystal structure by merging data mining with quantum mechanics. *Nature Mater.* **5**(8), 641–646 (2006).
108. T. Fix, S-L. Sahonta, V. Garcia, J.L. MacManus-Driscoll, and M.G. Blamire: Structural and Dielectric Properties of $SnTiO_3$, a putative ferroelectric. *Cryst. Growth Des.* **11**, 1422–1426 (2011).
109. A. Jain, G. Hautier, C.J. Moore, B. Kang, J. Lee, H. Chen, N. Twu, and G. Ceder: A computational investigation of $Li_9M_3(P_2O_7)_3(PO_4)_2$ (M = V, Mo) as cathodes for Li ion batteries. *J. Electrochem. Soc.* **159**(5), A622–A633 (2012).
110. Q. Kuang, J. Xu, Y. Zhao, X. Chen, and L. Chen: Layered monodiphosphate $Li_9V_3(P_2O_7)_3(PO_4)_2$: A novel cathode material for lithium-ion batteries. *Electrochim. Acta* **56**(5), 2201–2205 (2011).
111. H. Chen, G. Hautier, and G. Ceder: Synthesis, computed stability and crystal structure of a new family of inorganic compounds: Carbonophosphates. *J. Am. Chem. Soc.* **134**(48), 19619–19627 (2012).
112. G. Hautier, A. Jain, H. Chen, C. Moore, SP. Ong, and G. Ceder: Novel mixed polyanions lithium-ion battery cathode materials predicted by high-throughput ab initio computations. *J. Mater. Chem.* **21**, 17147–17153 (2011).
113. C. Jähne, C. Neef, C. Koo, H-P. Meyer, and R. Klingeler: A new $LiCoPO_4$ polymorph via low temperature synthesis. *J. Mater. Chem. A* **1**(8), 2856 (2013).
114. K. Snyder, B. Raguž, W. Hoffbauer, R. Glaum, H. Ehrenberg, and M. Herklotz: Lithium copper(I) orthophosphates $Li_{3-x}Cu_xPO_4$: Synthesis, crystal structures, and electrochemical properties. *Z. Anorg. Allg. Chem.* **640**(5), 944–951 (2014).
115. E. Mosymow, R. Glaum, and R.K. Kremer: Searching for "$LiCr^{II}PO_4$". *J. Solid State Chem.* **218**, 131–140 (2014).
116. L. Yang and G. Ceder: Data-mined similarity function between material compositions. *Phys. Rev. B: Condens. Matter Mater. Phys.* **88**, 224107 (2013).
117. M.W. Gaultois, A.O. Oliynyk, A. Mar, T.D. Sparks, G.J. Mulholland, and B. Meredig: A recommendation engine for suggesting unexpected thermoelectric chemistries. 7, (2015), 7arXiv: 1502.07635.
118. A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka: Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization, *Phys. Rev. Lett.* **115**(20), 205901 (2015).
119. H. Turner and D. Firth: Bradley-Terry models in R: The BradleyTerry2 Package. *J. Stat. Software* **48**(9), 1–21 (2012).
120. R.A. Bradley and M.E. Terry: Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).
121. J. Robertson and S.J. Clark: Limits to doping in oxides. *Phys. Rev. B* **83**(7), 075205 (2011).
122. D.O. Scanlon and G.W. Watson: On the possibility of p-type SnO_2. *J. Mater. Chem.* **22**(48), 25236 (2012).
123. A. Zunger: Practical doping principles. *Appl. Phys. Lett.* **83**(1), 57 (2003).
124. H. Kawazoe, M. Yasukawa, and H. Hyodo: P-type electrical conduction in transparent thin films of $CuAlO_2$. *Nature* **389**, 939–942 (1997).
125. S. Sheng, G. Fang, C. Li, S. Xu, and X. Zhao: p-type transparent conducting oxides. *Phys. Status Solidi A* **203**(8), 1891–1900 (2006).
126. A. Kudo, H. Yanagi, H. Hosono, and H. Kawazoe: $SrCu_2O_2$: A p-type conductive oxide with wide band gap. *Appl. Phys. Lett.* **73**(2), 220 (1998).
127. G. Trimarchi, H. Peng, J. Im, A. Freeman, V. Cloet, A. Raw, K. Poeppelmeier, K. Biswas, S. Lany, and A. Zunger: Using design principles to systematically plan the synthesis of hole-conducting transparent oxides: Cu_3VO_4 and Ag_3VO_4 as a case study. *Phys. Rev. B* **84**(16), 165116 (2011).
128. A. Walsh and J.L.F. Da Silva, S-H. Wei: Multi-component transparent conducting oxides: Progress in materials modelling. *J. Phys.: Condens. Matter* **23**(33), 334210 (2011).
129. G. Hautier, A. Miglio, G. Ceder, G-M. Rignanese, and X. Gonze: Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **4**, 2292 (2013).
130. H. Peng and S. Lany: Semiconducting transition-metal oxides based on d$5 cations: Theory for MnO and Fe_2O_3. *Phys. Rev. B: Condens. Matter Mater. Phys.* **85**(85), 201202 (2012).
131. S. Arlot and A. Celisse: A survey of cross-validation procedures for model selection. *Stat. Surveys* **4**, 40–79 (2010).

Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1557/jmr.2016.80.